Error-Flagging Support for Testing and Its Effect on Adaptation

Amruth N. Kumar

Ramapo College of New Jersey,
Mahwah, NJ 07430, USA
amruth@ramapo.edu

Abstract. The effect of providing error-flagging support during tests was studied in spring 2009 with two tutors on `while` and `for` loops. A partial crossover design was used for the study and mixed-factor ANOVA was used to analyze the students' score per problem, number of revised problems, effect of revision on score and time spent per problem, and finally, the effect of error-flagging on adaptation and learning. Students scored better on tests with rather than without error-flagging support. This can be attributed to the fact that students revised their answers on more problems when error-flagging feedback was provided. But, they did not necessarily revise more often per problem with error-flagging feedback. Students scored the same with revisions as without revisions, whether or not error-flagging support was provided. They spent less time per problem when error-flagging support was provided, whether or not they revised their answers. One explanation for this is that error-flagging may speed up the problem-solving process. Finally, if error-flagging feedback is provided during pre-test, students solve significantly fewer problems during the subsequent practice session which uses the outcome of the pre-test as the basis for adaptation. Therefore, providing error-flagging feedback during the pre-test improves adaptation. But, it does not result in greater learning - learning was not significantly different with versus without error-flagging feedback.

Keywords: Error-flagging, Testing, Adaptation, Evaluation.

1 Introduction

One mechanism proposed to build the student model needed for adaptation in tutors is pre-testing (e.g., [1, 5]). In order to be accurate, pre-tests must avoid both false-positives, when students can solve a problem correctly without knowing the underlying concept, and false-negatives, when they end up incorrectly solving a problem in spite of knowing the underlying concept. One approach used by instructional experts to minimize false positives is to design pre-test problems that require more than recall and recognition, e.g., problems with short-entry rather than multiple-choice answers. False negatives can occur when students incorrectly solve a problem because they misunderstand the instructions or the user interface, or second-guess themselves. The hypothesis of this paper is that providing error-flagging support, i.e., error-detection,

V. Aleven, J. Kay, and J. Mostow (Eds.): ITS 2010, Part I, LNCS 6094, pp. 359–368, 2010.

but not error-correction support during pre-test can improve student scores, possibly by minimizing false-negatives. If so, it would in turn improve adaptation that is based on the pre-test.

The effect of providing error-flagging feedback during testing has been studied with mixed results. Multiple studies of paper-and-pencil testing have reported lower performance due to increased anxiety (e.g., [3, 6]) or no difference (e.g., [12]) when feedback about the correctness of answers was provided. Studies with early Computer Assisted Instruction/Testing showed better performance with such feedback during testing than without (e.g., [2, 13]). Later studies with computer-based multiple-choice testing showed no relative advantage or performance gain from providing such feedback [11, 12]. In one of the most recent studies to our knowledge, researchers found that there was little difference among the types of feedback provided during testing with the ACT Programming Tutor [4]. In earlier preliminary studies using a tutor on arithmetic expression evaluation, we had found that error-flagging support helped students improve their test scores [8, 9].

Given the mixed nature of prior results, we revisited the issue of providing error-flagging feedback during testing. This study differs from many of the earlier studies in that the testing was done online; the problems were short-entry rather than multiple-choice in nature; and the outcome of the test was used for adaptation of problem-solving practice by a conflated software tutor.

1.1 Experimental Setup

In spring 2009, two problem-solving software tutors were used to evaluate the effect of providing error-flagging support during testing. The tutors were on two introductory programming concepts: `while` loops and `for` loops. The `while` loop tutor targeted 9 concepts and the `for` loop tutor targeted 10 concepts, such as zero-iteration execution, and dependent and independent nested loops. The tutors presented problems on these concepts, each problem containing a program whose output was to be determined by the student. The student entered the output free-hand (as opposed to selecting it from a menu of options).

Each software tutor went through the pre-test-practice-post-test protocol as follows:

- It first administered a pre-test to evaluate the prior knowledge of students and build the student model. The pre-test consisted of one problem per concept – 9 problems for the `while` loop tutor and 10 problems for the `for` loop tutor. Students were expected to attempt all the problems, although they had the option to discontinue the tutoring session at any time.
- Subsequently, it provided practice problems on only those concepts on which students had solved problems incorrectly during the pre-test [7];
- Finally, it administered post-test problems on only those concepts on which students had solved sufficient number of problems during practice as indicated by the student model.

The three stages were administered back-to-back without any break in between. The software tutors allowed 30 minutes for the three stages combined.

The evaluations conducted in spring 2009 were *in-vivo*. The software tutors were used by 365 students in the introductory programming course at 12 institutions, which were randomly assigned to one of two groups: A or B. Subjects, i.e., students accessed the tutors over the web, typically, after class. The software tutors remotely collected the data for analysis.

A partial cross-over design was used: students in group A served as control subjects on the `while` loop tutor and test subjects on the `for` loop tutor, while students in group B served as test subjects on the `while` loop tutor and control subjects on the `for` loop tutor. Both the groups worked with the `while` loop tutor before the `for` loop tutor. All else being equal, error-flagging feedback was provided during pre-test to students in the test group, but not the control group.

Error-flagging support was provided while the student was entering the answer to each problem, i.e., before the student submitted the answer. With error-flagging feedback, whenever the student entered each step in the answer to a problem, the step in the answer was displayed with red background if incorrect and green background if correct. When incorrect, no facility was provided for the student to find out why the step was incorrect, or how it could be corrected. Without error-flagging support, the steps in a student's answer were always displayed with white background. The online instructions presented to the students before using each tutor explained the significance of the background colors.

Whether or not the tutor provided error-flagging feedback, students had the option to revise their answer as often as necessary before submitting it. This included deleting or editing any step(s) in the answer. Once again, the instructions presented to the students before using each tutor explained the user interface facilities provided for revising an answer.

2 Results

For analysis, only those students were considered who had attempted most of the pre-test problems, i.e., at least 7 of the 9 problems on the `while` loop pre-test and at least 8 of the 10 problems on the `for` loop pre-test. In order to factor out the effect of the difference in the number of problems solved by students, the average score per problem was considered for analysis, which can range from 0 through 1, rather than the total score. Similarly, the average time spent per problem was considered rather than the total time spent on the pre-test.

Score Per Problem: A 2 X 2 mixed-factor ANOVA analysis of the score per problem was conducted with the topic (`while` versus `for` loop) as the repeated measure and the group (group A with error-flagging on `for` loop pre-test versus group B with error-flagging on `while` loop pre-test) as the between subjects factor.

A large significant interaction was found between topic and group [F(1,363) = 216.563, $p < 0.001$]. As shown in Table 1, both the groups scored better with error-flagging support than without: group A scored 0.624 on `while` loop pre-test without error-flagging support, and went on to score 0.871 on `for` loop pre-test with error-flagging support. The difference was statistically significant [t(289) = 18.578, p < 0.001]. Group B scored 0.842 on `while` loop pre-test with error-flagging support, and went on to score 0.667 on `for` loop pre-test without error-flagging support. The difference was statistically significant [t(74) = -7.565, p < 0.001]. *So, students scored better on tests with rather than without error-flagging support.*

Table 1. Average Pre-test Score with and Without Error-Flagging

	`While` loop pre-test	`for` loop pre-test
Without Error-Flagging	0.624	0.667
With Error-Flagging	0.842	0.871

Number of Revised Problems: Did the provision of error-flagging support result in subjects revising their answers on more problems? In order to answer this question, the pre-test problems solved by each student were grouped into those where the student revised the answer versus those where the student never revised the answer. The 2 X 2 mixed-factor ANOVA analysis was repeated on the number of problems on which subjects revised their answer, with topic as the repeated measure and group as the between subjects factor.

A significant and large interaction was observed between topic and group [F(1,369) = 399.836, p < 0.001]. Group A revised 0.990 problems on `while` loop pre-test when no error-flagging support was provided, and then went on to revise 5.956 problems on `for` loop pre-test when error-flagging support was provided, as shown in Table 2. The difference was statistically significant [t(293) = -28.686, p < 0.001]. Group B revised answers on 3.416 problems on `while` loop pre-test when error-flagging feedback was provided, but then, went on to revise only 1.00 problem on `for` loop pre-test when no error-flagging feedback was provided. The difference was statistically significant [t(76) = 8.376, p < 0.001]. *So, students revised their answers on more problems when error-flagging feedback was provided than when it was not.*

Table 2. Number of revised problems with and without Error-Flagging

	`while` loop pre-test	`for` loop pre-test
Without Error-Flagging	0.990	1.000
With Error-Flagging	3.416	5.956

Number of Revisions per Problem: Did the provision of error-flagging support result in subjects revising their answers more often per problem? A univariate analysis of the number of revisions per problem was conducted on `while` loop pre-test data with the problem number (1-9) and error-flagging (without versus with) as fixed factors. A significant main effect was found for problem [F(8,680) = 6.223, p < 0.001], indicating that the number of revisions per problem was not uniform across the board, but rather, depended on the problems. A significant main effect was found for error-flagging [F(1,680) = 18.98, p < 0.001]: revisions per problem was lower with error-flagging (1.566) than without (2.143). But, no significant interaction was found between problem and error-flagging [F(8,680) = 1.146, p = 0.33]. A similar analysis was conducted on `for` loop pre-test data. Once again, a significant main effect was found for problem [F(9,2096) = 2.202, p = 0.019]. But, no significant main effect was found for error-flagging [F(1,2096) = 0.137, p = 0.711]. The interaction between problem and error-flagging was marginally significant [F(9,2096) = 1.828, p = 0.059]: revisions per problem were more with error-flagging on some problems, but not others. *So, error-flagging feedback did not necessarily result in more revisions per problem.*

Effect of Revision on Score: In order to evaluate the effect of revision on pre-test score, a 2 X 2 X 2 mixed-factor ANOVA analysis was conducted of the average score per problem, with the topic (`while` versus `for`) and revision (without versus with revision) as within-subjects factors and group (group A with error-flagging on `for` loop pre-test versus group B with error-flagging on `while` loop pre-test) as between-subjects factor.

No significant main effect was found for revision [F(1,196) = 2.088, p = 0.15]. No significant interaction was found between revision and group [F(1,196) = 0.507, p = 0.477], topic and revision [F(1,196) = 0.423, p = 0.516] or topic, revision and group [F(1, 196) = 0.335, p = 0.564]. The score per problem on problems with and without revision for the two groups is listed in Table 3. None of the differences in the scores between revised and unrevised problems was statistically significant. No ceiling effect was observed in the scores either. *In other words, students scored the same with revisions as without revisions, whether or not error-flagging support was provided.*

Table 3. Score per problem with versus without revision: Group A got error-flagging feedback during `for` loop pre-test and Group B got error-flagging feedback during `while` loop pre-test

	`while` loop pre-test		`for` loop pre-test	
	No Revision	Revised	No Revision	Revised
Group A	0.606	0.667	0.865	0.883
Group B	0.878	0.892	0.672	0.684

Effect of Revision on Time Spent: In order to evaluate the effect of revisions on the time spent solving problems, a 2 X 2 X 2 mixed-factor ANOVA analysis of the average time spent per problem was conducted, with the topic (`while` versus `for`) and revision (without versus with) as within-subjects factors and group (group A with error-flagging on `for` loop pre-test versus group B with error-flagging on `while` loop pre-test) as between-subjects factor.

A significant interaction was found between topic and group [F(1,194) = 28.636, p < 0.001]. Group A spent 142.629 seconds per problem on `while` loop pre-test without error-flagging support and went on to spend 97.962 seconds per problem on `for` loop pre-test with error-flagging support, as shown in Table 4. Group B spent 99.467 seconds per problem on `while` loop pre-test with error-flagging support and went on to spend 114.915 seconds per problem on `for` loop pre-test without error-flagging support. *So, students spent less time per problem when error-flagging feedback was provided.*

Table 4. Time spent per problem with and without Error-Flagging

	`while` loop pre-test	`for` loop pre-test
Without Error-Flagging	142.629	114.915
With Error-Flagging	99.467	97.962

The interaction between topic and group was significant for both revised problems [F(1,221) = 9.848, p = 0.002] and unrevised problems [F(1,328) = 56.170, p < 0.001]. For revised problems, group A spent 172.084 seconds per problem without error-flagging feedback on `while` loop pre-test, followed by 112.781 seconds per problem with error-flagging feedback on `for` loop pre-test. Group B spent 125.168 seconds per problem with error-flagging feedback and 129.282 seconds per problem without error-flagging feedback as shown in Table 5. For unrevised problems, group A spent 109.780 seconds per problem without error-flagging support on `while` loop pre-test, followed by 82.866 seconds per problem with error-flagging support on `for` loop pre-test. Group B spent 79.063 seconds per problem with error-flagging support and 101.499 seconds per problem without error-flagging support as shown in Table 6. *So, whether or not students revised their answers, they solved problems faster with error-flagging.*

Table 5. Time spent per problem with and without Error-Flagging on revised problems

Revised Problems	`while` loop pre-test	`for` loop pre-test
Without Error-Flagging	172.084	129.282
With Error-Flagging	125.168	112.781
Significance	t(246) =-3.028, p = 0.003	t(328) = 2.125, p = 0.034

Table 6. Time spent per problem with and without Error-Flagging on unrevised problems

Unrevised Problems	`while` loop pre-test	`for` loop pre-test
Without Error-Flagging	109.780	101.499
With Error-Flagging	79.063	82.866
Significance	t(368) = -5.301, p < 0.001	t(329) = 3.945, p < 0.001

No significant interaction was observed between revision and group [F(1,194) = 0.075, p = 0.784]. No significant interaction was found between topic, revision and group [F(1,194) = 0.104, p = 0.748]. A significant main effect was found for revision [F(1,194) = 59.488, p < 0.001] – students spent an average of 94.183 seconds per problem when they did not revise their answer, and 133.304 seconds per problem when they did revise their answer. *So, students spent more time per problem when they revised their answer than when they did not.* This is to be expected since revising an answer involves undoing and redoing one or more steps in the answer.

Effect of Error-Flagging on Subsequent Adaptation: Since students score better on pre-test with error-flagging support, do they solve fewer problems during the subsequent practice that uses the results of the pre-test as the basis for adaptation? The number of problems solved during adaptive practice was analyzed using 2 X 2 mixed factor ANOVA with topic (while versus for) as within-subjects factor and group (group A with error-flagging on for loop pre-test versus group B with error-flagging on while loop pre-test) as between-subjects factor. No significant main effect was observed for topic [F(1,279) = 0.058, p = 0.810]. (The smaller N is due to the fact that some students did not solve any practice problems since they answered all the pre-test problems correctly). A significant main effect was observed for group [F(1,279) = 5.722, p = 0.017]. This is due to the fact that one group solved more problems with and without error-flagging than the other group, as shown in Table 7.

Significant interaction was observed between topic and group [F(1,279) = 25.227, p < 0.001]. As shown in Table 7, both the groups solved fewer problems during adaptive practice when error-flagging support was provided during the preceding pre-test than when it was not provided: group A solved 10.897 practice problems on while loop, given no error-flagging support during pre-test, and went on to solve 8.268 practice problems on for loop, given error-flagging support during pre-test. The difference was statistically significant [t(223) = -5.458, p < 0.001]. Group B solved 10.070 practice problems on while loop, given error-flagging support during pre-test, and went on to solve 12.965 practice problems on for loop, given no error-flagging support during pre-test. The difference was statistically significant [t(56) = 2.674, p = 0.01]. *So, students solved significantly fewer problems during adaptive practice when error-flagging support was provided during the preceding pre-test than when it was not.*

Table 7. Problems Solved During Adaptive Practice with and without Error-Flagging Feedback Provided during the Preceding Pre-test

	while loop practice	for loop practice
Pre-test without Error-Flagging	10.897	12.965
Pre-test with Error-Flagging	10.070	8.268

Learning: After the adaptive practice, students answered a post-test on only the concepts on which they had solved sufficient number of problems during practice as indicated by the student model. The learning of students was measured in terms of their pre-test and post-test scores on only those concepts (henceforth referred to as learned concepts) on which they had solved problems during all three stages: pre-test, adaptive practice and post-test. So, analysis of learning excludes the records of students who solved all the problems correctly on the pre-test, and hence, did not solve any problems during practice or post-test; and the records of students who ran out of time either during practice or post-test, since they were allowed 30 minutes for the three stages combined. In other words, only those students were considered for this analysis who learned at least one concept using the tutor, and for these students, data from only the learned concepts was used.

A 2 X 2 mixed-factor ANOVA analysis was conducted of the score per problem on the learned concepts, with pre-post as repeated measure and error-flagging (with versus without) as between-subjects factor. On the while loop tutor, a significant main effect was observed for pre-post [F(1,98) = 546.021, p < 0.001]: student scores improved from 0.262 on the pre-test to 0.933 on the post-test. A marginally significant main effect was observed for error-flagging [F(1,98) = 3.800, p = 0.054]: 0.630 with error-flagging and 0.565 without error-flagging. No significant interaction was observed between pre-post and error-flagging [F(1,98) = 0.404, p = 0.526]: the group that got error-flagging feedback scored slightly better than the group that did not, on both pre-test and post-test.

On the `for` loop tutor, once again, a significant main effect was observed for pre-post [F(1,138) = 508.976, p < 0.001]: student scores improved from 0.279 on the pre-test to 0.901 on the post-test. No significant main effect was observed for error-flagging [F(1,138) = 2.731, p = 0.101] and no significant interaction was observed between pre-post and error-flagging [F(1,138) = 1.258, p = 0.264]. *So, student learning was not significantly different with versus without error-flagging.*

3 Discussion

An empirically-driven study was conducted to see whether providing error-flagging support, i.e., error-detection, but not error-correction support during tests could improve student scores, possibly by minimizing false-negatives. Since error-flagging support could promote revision of incorrect answers by students, the number of problems on which students revised their answers, the number of revisions per problem, and the score and time per problem with versus without error-flagging and with versus without revisions were all studied. The results of the study could have implications for adaptation in tutors, and for online testing in general.

Students scored better on tests with rather than without error-flagging support. This can be attributed to the fact that they revised their answers on more problems when error-flagging feedback was provided. These revisions could have eliminated false negatives, resulting in higher scores. However, students did not necessarily revise more often per problem with error-flagging than without – they actually revised less often per problem with error-flagging on the `while` loop pre-test. This counter-intuitive result suggests that error-flagging does not promote indiscriminate revisions of answers by students, i.e., students do not necessarily abuse error-flagging support to guess the correct answer through trial and error. Moreover, students scored the same per problem with revisions as without revisions, whether or not error-flagging support was provided.

Students spent more time per problem when they revised their answer than when they did not. They spent less time per problem when error-flagging feedback was provided. For the `while` loop pre-test, this can be explained based on the fact that students revised their answers significantly less often per problem with error-flagging (1.566) than without (2.143) [F(1,680) = 18.98, p < 0.001]. One explanation for why students revised less often per problem with error-flagging is that given confirmation of the correctness of the answer so far, students may forgo some *false negatives*, revisions that they would have otherwise made, instances where they would have second-guessed a correct answer that they had already entered.

But, this does not explain why students spent less time per problem with error-flagging on `for` loop pre-test, wherein revisions per problem was not significantly different without (2.085) and with error-flagging (1.983) [F(1,2096) = 0.137, p = 0.711]. A 2 X 2 ANOVA of the time spent per problem during `for` loop pre-test with revision (with versus without) as within-subjects factor and error-flagging (with versus without) as between-subjects factor found a significant main effect for revision [F(1,288) = 37.792, p < 0.001] and error-flagging [F(1,288) = 9.343, p = 0.002], but no significant interaction between revision and error-flagging [F(1,288) = 0.284, p = 0.594]. So, students spent less time per problem with error-flagging (97.088 seconds) than without error-flagging (114.666 seconds); they spent less time without revision (92.921 seconds) than with revision (118.833 seconds); yet, they did not revise any significantly less per problem with error-flagging than without. This suggests that with error-flagging feedback, either students revised their answer more quickly, or saved time that was attributable to the unrevised parts of the answer, or both.

Indeed, whether or not students revised their answers, they solved problems faster with error-flagging than without. One possible explanation is that students proceed more quickly to revise their answer when it is marked incorrect, and proceed more quickly to the next step in the answer when it is marked correct by error-flagging feedback. In both cases, they save on the time they would have optionally spent reconsidering the correctness of the answer they had just entered. In other words, error-flagging feedback may have the effect of speeding up the problem-solving process. Testing this hypothesis quantitatively is part of our future work.

If error-flagging feedback is provided during pre-test, students solve significantly fewer problems during the subsequent practice session which uses the outcome of the pre-test as the basis for adaptation. Therefore, providing error-flagging feedback during the pre-test improves adaptation. Any adaptive system that uses a pre-test to build the initial student model would benefit from providing error-flagging support during the pre-test. However, error-flagging does not result in greater learning - learning was not significantly different with versus without error-flagging.

To further generalize this result, given that it is logistically easier to provide error-flagging support during online tests (as opposed to pen-and-paper tests), and such support helps students score better and answer faster even when the test items are not multiple-choice in nature, provision of error-flagging support should be considered in all online tests.

Acknowledgments. Partial support for this work was provided by the National Science Foundation under grant DUE-0817187**.**

References

1. Aimeur, E., Brassard, G., Dufort, H., Gambs, S.: CLARISSE: A Machine Learning Tool to Initialize Student Models. In: Cerri, S.A., Gouardéres, G., Paraguaçu, F. (eds.) ITS 2002. LNCS, vol. 2363, pp. 718–728. Springer, Heidelberg (2002)
2. Anderson, R.C., Kulhavy, R.W., Andre, T.: Feedback procedures in programmed instruction. J. Educational Psychology 62, 148–156 (1971)
3. Bierbaum, W.B.: Immediate knowledge of performance on multiple-choice tests. J. Programmed Instruction 3, 19–23 (1965)
4. Corbett, A.T., Anderson, J.R.: Locus of feedback control in computer-based tutoring: impact on learning rate, achievement and attitudes. In: Proc. SIGCHI Conference on Human Factors in Computing Systems, pp. 245–252 (2001)

5. Czarkowski, M., Kay, J.: Challenges of Scrutable Adaptivity. In: Proc. of AI-ED 2003, pp. 404–406. IOS Press, Amsterdam (2003)
6. Gilmer, J.S.: The Effects of Immediate Feedback Versus Traditional No-Feedback in a Testing Situation. In: Proc. Annual Meeting of the American Educational Research Association, April 1979, pp. 8–12 (1979)
7. Kumar, A.N.: A Scalable Solution for Adaptive Problem Sequencing and its Evaluation. In: Wade, V.P., Ashman, H., Smyth, B. (eds.) AH 2006. LNCS, vol. 4018, pp. 161–171. Springer, Heidelberg (2006)
8. Kumar, A.N.: The Effect of Providing Error-Flagging Support during Testing. In: Woolf, B.P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) ITS 2008. LNCS, vol. 5091, pp. 799–802. Springer, Heidelberg (2008)
9. Kumar, A.N., Rutigliano, P.: The Effects of Error-Flagging in a Tutor on Expression Evaluation. In: 13th International Conference on Artificial Intelligence in Education (AI-ED 2007), pp. 599–601 (2007)
10. Montor, K.: Effect of using a self scoring answer sheet on knowledge retention. J. Educational Research 63, 435–437 (1970)
11. Plake, B.S.: Effects of Informed Item Selection on Test Performance and Anxiety for Examinees Administered a Self-Adapted Test. Educational and Psychological Measurement 55(5), 736–742 (1995)
12. Shermis, M.D., Mzumara, H.R., Bublitz, S.T.: On Test and Computer Anxiety: Test Performance Under CAT and SAT Conditions. J. Education Computing Research 24(10), 57–75 (2001)
13. Tait, K., Hartley, J.R., Anderson, R.C.: Feedback procedures in computer-assisted arithmetic instruction. British Journal of Educational Psychology 43, 161–171 (1973)